Naive_Bayes

David Suffolk

5/2/2020

Naive Bayes

Import Data

```
library(naivebayes)
```

```
## naivebayes 0.9.7 loaded
```

```
data <- read.csv("./Project 02/nwCrow_bloodParasites_alaska_smith_2007_2008/nwCrow_sampling_alaska_smith
data_2 <- read.csv("./Project 02/nwCrow_bloodParasites_alaska_smith_2007_2008/nwCrow_bloodParasites_alas
de <- merge(data, data_2, by=0, all=TRUE)
head(de)
```

```
##   Row.names Field.ID      DATE  LOC   LAT    LONG SEX AGE AKD TARSUS WING MASS
## 1         1    75001 3/20/2007 SEWA 60.11 -149.44   1   1   1   55.8  283  448
## 2        10    75010 3/22/2007 KENA 60.55 -151.23   2   1   0   48.6  271  390
## 3       100    75100 3/12/2008 VALD 61.12 -146.35   2   1   0   44.6  264  317
## 4       101    86701 3/12/2008 VALD 61.12 -146.35   2   1   0   47.1  269  343
## 5       102    86702 3/12/2008 VALD 61.12 -146.35   2   2   0   52.2  291  415
## 6       103    86703 3/12/2008 VALD 61.12 -146.35   1   2   0   47.0  266  325
##   Extraction.. LEUC1 LEUC2 HAEM1 HAEM2 PLAS1 PLAS2 Leuc_GenBank_Accession
## 1      NOCR001     0     0     0     0     0     0
## 2      NOCR010     0     0     0     0     0     0
## 3      NOCR100     1     1     0     0     0     0               MG765394
## 4      NOCR101     0     0     0     0     0     0
## 5      NOCR102     1     0     0     0     0     0               MG765394
## 6      NOCR103     1     1     0     0     0     0               MG765394
##   Haem_GenBank_Accession Plas_GenBank_Accession
## 1
## 2
## 3
## 4
## 5
## 6
```

One Hot Encoding

```
for(unique_value in unique(de$LOC)){


de[paste("LOC", unique_value, sep = ".")] <- ifelse(de$LOC == unique_value, 1, 0)

}
head(de)
```

```
##   Row.names Field.ID      DATE  LOC   LAT    LONG SEX AGE AKD TARSUS WING MASS
## 1         1    75001 3/20/2007 SEWA 60.11 -149.44   1   1   1   55.8  283  448
## 2        10    75010 3/22/2007 KENA 60.55 -151.23   2   1   0   48.6  271  390
## 3       100    75100 3/12/2008 VALD 61.12 -146.35   2   1   0   44.6  264  317
## 4       101    86701 3/12/2008 VALD 61.12 -146.35   2   1   0   47.1  269  343
## 5       102    86702 3/12/2008 VALD 61.12 -146.35   2   2   0   52.2  291  415
## 6       103    86703 3/12/2008 VALD 61.12 -146.35   1   2   0   47.0  266  325
##   Extraction.. LEUC1 LEUC2 HAEM1 HAEM2 PLAS1 PLAS2 Leuc_GenBank_Accession
## 1      NOCR001     0     0     0     0     0     0
## 2      NOCR010     0     0     0     0     0     0
## 3      NOCR100     1     1     0     0     0     0               MG765394
## 4      NOCR101     0     0     0     0     0     0
## 5      NOCR102     1     0     0     0     0     0               MG765394
## 6      NOCR103     1     1     0     0     0     0               MG765394
##   Haem_GenBank_Accession Plas_GenBank_Accession LOC.SEWA LOC.KENA LOC.VALD
## 1                                                      1        0        0
## 2                                                      0        1        0
## 3                                                      0        0        1
## 4                                                      0        0        1
## 5                                                      0        0        1
## 6                                                      0        0        1
##   LOC.HAIN LOC.JUNE LOC.HOME
## 1        0        0        0
## 2        0        0        0
## 3        0        0        0
## 4        0        0        0
## 5        0        0        0
## 6        0        0        0
```

Filter Columns

```
de <- de[,c(7,8,9,10,11,12,14,16,18,23,24,25,26,27,28)]
head(de)
```

```
##   SEX AGE AKD TARSUS WING MASS LEUC1 HAEM1 PLAS1 LOC.SEWA LOC.KENA LOC.VALD
## 1   1   1   1   55.8  283  448     0     0     0        1        0        0
## 2   2   1   0   48.6  271  390     0     0     0        0        1        0
## 3   2   1   0   44.6  264  317     1     0     0        0        0        1
## 4   2   1   0   47.1  269  343     0     0     0        0        0        1
## 5   2   2   0   52.2  291  415     1     0     0        0        0        1
## 6   1   2   0   47.0  266  325     1     0     0        0        0        1
##   LOC.HAIN LOC.JUNE LOC.HOME
## 1        0        0        0
```

```
## 2        0        0        0
## 3        0        0        0
## 4        0        0        0
## 5        0        0        0
## 6        0        0        0
```

Factoring

```
de$SEX <- as.factor(de$SEX)
de$AGE <- as.factor(de$AGE)
de$AKD <- as.factor(de$AKD)
de$TARSUS <- as.factor(de$TARSUS)
de$WING <- as.factor(de$WING)
de$MASS <- as.factor(de$MASS)
de$LEUC1 <- as.factor(de$LEUC1)
de$HAEM1 <- as.factor(de$HAEM1)
de$PLAS1 <- as.factor(de$PLAS1)
de$LOC.SEWA <- as.factor(de$LOC.SEWA)
de$LOC.KENA <- as.factor(de$LOC.KENA)
de$LOC.VALD <- as.factor(de$LOC.VALD)
de$LOC.HAIN <- as.factor(de$LOC.HAIN)
de$LOC.JUNE <- as.factor(de$LOC.JUNE)
de$LOC.HAIN <- as.factor(de$LOC.HAIN)
```

Data Partition

```
set.seed(1234)
ind <- sample(2, nrow(de), replace = T, prob = c(0.8,0.2))
train <- de[ind == 1,]
test <- de[ind == 2,]
```

AKD

Model

```
model <- naive_bayes(AKD ~ ., data = train, laplace=1)
```

Predict

```
p <- predict(model, train, type = 'prob')
```

```
## Warning: predict.naive_bayes(): more features in the newdata are provided as
## there are probability tables in the object. Calculation is performed based on
## features to be found in the tables.
```

Confusion Matrix

Train

```
p1 <- predict(model, train)
```

```
## Warning: predict.naive_bayes(): more features in the newdata are provided as
## there are probability tables in the object. Calculation is performed based on
## features to be found in the tables.
```

```
tab1 <- table(p1, train$AKD)
tab1
```

```
##
## p1    0   1
##   0 138   5
##   1   1  10
```

```
# Misclassification
incorrect <- 1 - sum(diag(tab1)) / sum(tab1)
correct<- 100*(1 - incorrect)
correct
```

```
## [1] 96.1039
```

Test

```
p2 <- predict(model, test)
```

```
## Warning: predict.naive_bayes(): more features in the newdata are provided as
## there are probability tables in the object. Calculation is performed based on
## features to be found in the tables.
```

```
tab2 <- table(p2, test$AKD)
tab2
```

```
##
## p2   0  1
##   0 28  4
##   1  0  0
```

```
# Misclassification
incorrect <- 1 - sum(diag(tab2)) / sum(tab2)
correct<- 100*(1 - incorrect)
correct
```

```
## [1] 87.5
```

LEUC1

Model

```
model <- naive_bayes(LEUC1 ~ ., data = train, laplace=1)
```

Predict

```
p <- predict(model, train, type = 'prob')
```

```
## Warning: predict.naive_bayes(): more features in the newdata are provided as
## there are probability tables in the object. Calculation is performed based on
## features to be found in the tables.
```

Confusion Matrix

Train

```
p1 <- predict(model, train)
```

```
## Warning: predict.naive_bayes(): more features in the newdata are provided as
## there are probability tables in the object. Calculation is performed based on
## features to be found in the tables.
```

```
tab1 <- table(p1, train$LEUC1)
tab1
```

```
##
## p1   0  1
##   0 69  9
##   1 10 66
```

```
# Misclassification
incorrect <- 1 - sum(diag(tab1)) / sum(tab1)
correct<- 100*(1 - incorrect)
correct
```

```
## [1] 87.66234
```

Test

```
p2 <- predict(model, test)
```

```
## Warning: predict.naive_bayes(): more features in the newdata are provided as
## there are probability tables in the object. Calculation is performed based on
## features to be found in the tables.
```

```
tab2 <- table(p2, test$LEUC1)
tab2
```

```
##
## p2  0 1
##   0 8 8
##   1 8 8
```

```
# Misclassification
incorrect <- 1 - sum(diag(tab2)) / sum(tab2)
correct<- 100*(1 - incorrect)
correct
```

```
## [1] 50
```

HAEM1

Model

```
model <- naive_bayes(HAEM1 ~ ., data = train, laplace=1)
```

Predict

```
p <- predict(model, train, type = 'prob')
```

```
## Warning: predict.naive_bayes(): more features in the newdata are provided as
## there are probability tables in the object. Calculation is performed based on
## features to be found in the tables.
```

Confusion Matrix

Train

```
p1 <- predict(model, train)
```

```
## Warning: predict.naive_bayes(): more features in the newdata are provided as
## there are probability tables in the object. Calculation is performed based on
## features to be found in the tables.
```

```
tab1 <- table(p1, train$HAEM1)
tab1
```

```
##
## p1    0   1
##   0 103   4
##   1   9  38
```

```r
# Misclassification
incorrect <- 1 - sum(diag(tab1)) / sum(tab1)
correct<- 100*(1 - incorrect)
correct
```

```
## [1] 91.55844
```

Test

```r
p2 <- predict(model, test)
```

```
## Warning: predict.naive_bayes(): more features in the newdata are provided as
## there are probability tables in the object. Calculation is performed based on
## features to be found in the tables.
```

```r
tab2 <- table(p2, test$HAEM1)
tab2
```

```
##
## p2   0  1
##   0 16  4
##   1  4  8
```

```r
# Misclassification
incorrect <- 1 - sum(diag(tab2)) / sum(tab2)
correct<- 100*(1 - incorrect)
correct
```

```
## [1] 75
```

PLAS1

Model

```r
model <- naive_bayes(PLAS1 ~ ., data = train, laplace=1)
```

Predict

```r
p <- predict(model, train, type = 'prob')
```

```
## Warning: predict.naive_bayes(): more features in the newdata are provided as
## there are probability tables in the object. Calculation is performed based on
## features to be found in the tables.
```

Confusion Matrix

Train

```
p1 <- predict(model, train)
```

```
## Warning: predict.naive_bayes(): more features in the newdata are provided as
## there are probability tables in the object. Calculation is performed based on
## features to be found in the tables.
```

```
tab1 <- table(p1, train$PLAS1)
tab1
```

```
##
## p1    0   1
##   0 138   4
##   1   3   9
```

```
# Misclassification
incorrect <- 1 - sum(diag(tab1)) / sum(tab1)
correct<- 100*(1 - incorrect)
correct
```

```
## [1] 95.45455
```

Test

```
p2 <- predict(model, test)
```

```
## Warning: predict.naive_bayes(): more features in the newdata are provided as
## there are probability tables in the object. Calculation is performed based on
## features to be found in the tables.
```

```
tab2 <- table(p2, test$PLAS1)
tab2
```

```
##
## p2   0  1
##   0 25  5
##   1  2  0
```

```
# Misclassification
incorrect <- 1 - sum(diag(tab2)) / sum(tab2)
correct<- 100*(1 - incorrect)
correct
```

```
## [1] 78.125
```